The Use of Fuzzy Linear Regression Models for Tumor Size in Colorectal Cancer in Hospital of Malaysia

Muhammad Ammar Shafi
Faculty of Science, Technology and Human Development
University Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor Malaysia

Mohd Saifullah Rusiman
Faculty of Science, Technology and Human Development
University Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor Malaysia

Copyright © 2015 Muhammad Ammar Shafi and Mohd Saifullah Rusiman. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract
Regression analysis has become popular among several fields of research and standard tools in analysing data. This structure was represented by four commonly statistical models such as multiple linear regression, fuzzy linear regression (Tanaka, 1982), fuzzy linear regression (Ni, 2005) and extended fuzzy linear regression by benchmarking models under fuzziness (Chung, 2012). Colorectal cancer (CRC) was applied on CRC cases in Malaysia. The CRC patients’ quality of life in order to detect the CRC at an early stage is still very poor, the programmes are mainly ad-hoc and not implemented as a national wide programme. This study aims to determine the best model to measure the tumor size at hospitals using mean square error and root mean square error. Secondary data was used where 180 patients having colorectal cancer and receiving treatment in hospitals was recorded by nurses and doctors. Based on the results, fuzzy linear regression (Ni, 2005) is the best model to predict the tumor size developed by patients after receiving treatment in hospital.
Keywords: Fuzzy Linear Regression, extended fuzzy linear regression by benchmarking models under fuzziness, multiple linear regression, mean square error, root mean square error

1. Introduction

Regression analysis has become one of the standard tools in analysing data. Its popularity is due to several reasons. The mathematical equation gained from its analysis could explain relationship between the dependent and independent variables. It provides much explanatory power, especially due to its multivariate nature. It is widely available in computer packages and easy to interpret. It has been widely used in applied sciences, economic, engineering, computer, social sciences and other fields (Agresti, 1996).

Nonlinear modelling is of interest to many researchers in modelling statistics, rather than linear modelling. The functional form obtained should be approximately near to the real data. If the functional form is far away from the real data, its mean the estimation is inconsistent, bias and so on. However, other difficulties may arise with a non-linear approach (Rousseeuw et. al, 2004). Nowadays, there are many models resulting from the regression analysis such as multiple regression, quadratic regression, cubic regression, logit model, probit model, exponential model, growth model, neural network regression and fuzzy regression. Statistic methods are suitable for use in the medical area. Other researchers in Malaysia conducted research in medical statistics such as mortality of patients in Intensive Care Unit (ICU) (Muhammad Ammar et. al, 2014; Mohd Saifullah et. al, 2012).

Colorectal cancer is cancer of the colon and rectum. The colon and rectum are two parts in the human body that play important roles in digesting food and producing waste. Colorectal cancer is one of the most common malignancies in the world (Malaysian Oncological Society, 2007). According to World Health Organization in 2012, colorectal cancer is the fourth leading cause of death among cancer sufferers. This cancer is a rising threat in many countries especially in the Asian Region including Malaysia.

Colorectal cancer is the third leading cause of cancer deaths in Malaysia. World Health Organization in 2012 stated lung cancer has the highest death rate caused by cancer at 17.93%. It is followed by breast cancers at 15.83% and colorectal cancer is third in the list at 13.10% rates of cancer deaths. Data from the Ministry of Health of Malaysia in 1995 indicate an increase in colorectal cancer admission rates from 8.1% to 11.9%. Recent studies have shown increasing colorectal cancer (CRC) cases in Asian population (Wendy, 2008; Radzi, 2008). According to the Second Paper of the National Cancer Registry (2003), 14.2% of male in Malaysia suffer from colorectal cancer while for female, it is 10.1%. It can be concluded that cancer is more common among men than women. CRC comprises of four stages which ranging from an early stage, second stage, third stage and final stage. Stage I refers to when the cancer is confined to the inner lining of the colon or rectum, stage II indicates that the cancer has spread
through the wall of the colon or rectum, stage III is when the cancer spreads to
nearby lymph nodes and lastly stage IV whereby the cancer spreads to distant
parts of the body, such as the liver or lungs (Malaysian Oncological Society,
2007).

There are twenty five variables for independent variable and one variable
for dependent variable in this study. Twenty five variables includes gender,
ethnic, age, icd10, TNM (Tumor, Nodes, Metastasis) staging, family history of
colon cancer, diabetes mellitus, crohn's disease, ulcerative colitis, polyp, history
of cancer(s), endometrial, gastric, small bowel, hepatobiliary, urinary tract,
ovarian, other cancer, intestinal obstruction, colorectal, weight loss, diarrhoea,
aemia, blood stool and abdominal pain. Finally, the dependent variable is
tumour size (mm). This study aims to determine the best model to measure the
tumor size at general hospital at Kuala Lumpur. The twenty five independent
variables also were analysed to look the relationship toward the tumor size.

2. Methodology

The statistical software SPSS 20, Excel 2010 and Matlab 2008 were used to
analyse the data. Exploratory data analysis was used to explore the behaviour of
the data. Some explorations on the demographic characteristic were applied. For
categorical variables, data was presented in frequency and percentage distribution
while for continuous variable, descriptive statistics measured by mean or average,
standard deviation and correlation of variables were applied. The parameter and
model were stated as:

Multiple Linear Regression
Regression analysis was developed by Sir Francis Galton in 19th century. Galton
studied the relation between heights of parents and children and he noted that
the heights of children of both tall and short parents appeared to “revert” or
“regress” to the mean group. He considered this tendency to be a regression to
“mediocrity”. He developed a mathematical description of this regression
tendency. The term regression persists to describe statistical relations between
variables (Kutner, 2004).

\[
Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... \beta_k X_{ik} + \epsilon_i(\beta), \quad i=1,...,N
\]

(3.1)

or \( Y = X\beta + \epsilon \)  \hspace{1cm} (3.2)

The function for least squares method is,

\[
S(\beta_0, \beta_1, \beta_2, ..., \beta_k) = S(\beta) = \sum_{j=1}^{d} \epsilon_j^2 \quad \text{or} \quad \epsilon^T \epsilon
\]

From (3.1), \( \epsilon(\beta) = Y - X\beta \)

Then, \( S(\beta) = (Y - X\beta)^T(Y - X\beta) \)

\[= Y^T Y - 2\beta^T X^T Y + \beta^T X^T X \beta \]
To minimize $S(\beta)$, we have to differentiate $S(\beta)$ with respect to $\beta$ where 
\[
\frac{\partial S}{\partial \beta} |_{\hat{\beta}} = 0
\]

Hence, the least squares estimator is,
\[
\hat{\beta} = (X^T X)^{-1} X^T Y
\]

The values fit by the equation $\beta_0 + \beta_1 x_i + ... + \beta_k x_k$ are denoted $i$, and the residuals $\epsilon_i$ are equal to $y_i - \hat{y}_i$, the difference between the observed and fitted values.

**Fuzzy Linear Regression (Tanaka 1982)**

Hideo Tanaka (1982) is the researcher who first proposed the fuzzy linear regression. Fuzzy model existed where human estimation and some systems is influential and should deal with fuzzy structure. In order to estimate the fuzzy parameters $A^*_e = (\alpha_e, \zeta_e)$ the following linear programming problem should be solved.

\[
\begin{align*}
\min_{\alpha, \zeta} & = \zeta_1 + ... + \zeta_g \\
\text{subject to} & \quad \zeta \geq 0 \\
\alpha^T x_e + (1-H) \sum_{f} \zeta_f | x_{ef} | & \geq y_e + (1-H) \epsilon_e \\
-\alpha^T x_e + (1-H) \sum_{f} \zeta_f | x_{ef} | & \geq -y_e + (1-H) \epsilon_e
\end{align*}
\]

We can obtain the best fitted model for the given data by solving the conventional linear programming problem in (3.3).

The fuzzy linear regression model (FLRM) can be stated as,
\[
Y = A_0(\alpha_0, \zeta_0) + A_1(\alpha_1, \zeta_1) x_1 + ... + A_g(\alpha_g, \zeta_g) x_g
\]

**Fuzzy Linear Regression (Ni, 2005)**

This model has been approached by Yongshen Ni on 2005 and this model is an extension approach from fuzzy linear regression model by Hideo Tanaka on 1982. Ni (2005) proposed the fuzzy linear regression model as,
Use of fuzzy linear regression models for tumor size

\[ \min_{\alpha} = \zeta_1 + \ldots + \zeta_g \]

Subject to \( \zeta \geq 0 \) and

\[ \alpha^T x_e + (1 - H) \sum_f \zeta_f | x_{ef} | \geq y_e \]

\[ -\alpha^T x_e + (1 - H) \sum_f c_f | x_{ef} | \geq -y_e \]  \hspace{1cm} (3.5)

**Extended FLR by Benchmarking Models under Fuzziness (Chung, 2012)**

William Chung proposed the Extended FLR by benchmarking models under fuzziness as,

\[ \min \frac{x_{jk} - \bar{x}_k}{s_k} \leq d_k \leq \max \frac{x_{jk} - \bar{x}_k}{s_k} \]  \hspace{1cm} (3.6)

where

- \( x_{jk} \) is the sample value
- \( \bar{x}_k \) is the sample average
- \( s_k \) is the standard deviation
- \( d_k \) is the inequality either negative values or positive values

subject to equation:

\[ \hat{Y}(x) = A_0 + A_1 \left( \frac{x_1 - \bar{x}_1}{s_1} \right) + A_2 \left( \frac{x_2 - \bar{x}_2}{s_2} \right) + \ldots + A_n \left( \frac{x_n - \bar{x}_n}{s_n} \right) \]  \hspace{1cm} (3.7)

3. **Results**

Demographically, the majority showed that most colorectal cancer of patients in general hospital at Kuala Lumpur was Malay and male. Most of them were in the average of 61 years old and cancer colon for most patients is in the colon side.

3.1. **Multiple Linear Regression**

Firstly, assumptions of multiple linear regression should be fulfilled before the data is analyzed. This assumes constant variance, normality and multicollinearity. All the assumptions were satisfied and hence, the results will be trustworthy.

Multiple linear regression is one of the common models used by statistician especially in medical health. This model was used to study and analyze twenty five predictor variables and to further detecting colorectal cancer stage. After the analysis was done; only eleven predictor variables were significance for
colorectal cancer. The significant variables were age at diagnosis, icd10 site, TNM staging, family history, crohn’s disease, history of cancer, gastric, ovarian, intestinal obstruction, anaemia and abdominal. The significance of variables was measured by p-value which must be less than 0.05.

Other than that, ANOVA analysis had been conducted. The results show that the mean square error term is 129.558 and RMSE is 11.3826. The p-value for the F test statistic was less than 0.05, indicating strong evidence of alternative hypothesis against the null hypothesis. The squared multiple correlation $R^2 = \frac{SSR}{SST} = \frac{21396.590}{41348.550} = 0.517$, indicate that 51.7% of the variability in the tumour size variable is explained by the twenty five independent variables.

All the significant variables influence the colorectal cancer effects. The estimated multiple linear regression models for symptoms and factors of colorectal cancer model are as follow:

$\hat{Y} = 76.056 + 0.421 \text{ age} + 3.459 \text{ icd10} + 0.961 \text{ TNM Staging} - 16.738 \text{ family history} + 5.035 \text{ Crohn’s disease} + 5.557 \text{ history of cancer} - 6.517 \text{ gastric} + 12.865 \text{ ovarian} - 4.350 \text{ intestinal obstruction} - 7.943 \text{ anaemia} - 3.994 \text{ abdominal}.$

3.2. Fuzzy Linear Regression (Tanaka 1982)
In this study, fuzzy linear regression by Tanaka (1982) was used to model tumor size faced by patients. The evaluation of fuzzy linear regression by Tanaka (1982) was based on mean square error and root mean square error. Mean square error for this model was 142.794 and root mean square error was 11.950.

The estimated fuzzy linear regression (Tanaka 1982) model for colorectal cancer of patients is as follow:

$\hat{Y} = -4.550 + (6.417, 0.426) \text{ gender} - (0.008, 0) \text{ age} + (5.545, -0.589) \text{ ethnic} + (8.794, 0.539) \text{ icd10} + (2.675, 0.473) \text{ TNM Staging} + (16.992, 2.932) \text{ family history} + (0.320, 1.197) \text{ diabetes mellitus} - (2.810, -0.745) \text{ Crohn’s disease} - (2.395, -1.611) \text{ ulcerative colitis} + (3.291, -1.013) \text{ polyp} - (3.980, 0.117) \text{ history of cancer} + (1.127, 0.041) \text{ endometrial} + (8.001, -0.465) \text{ gastric} + (6.299, 0.661) \text{ small bowel} + (1.972, -1.220) \text{ hepatobiliary} - (2.859, 0.659) \text{ urinary tract} - (11.855, -1.157) \text{ ovarian} + (0.499, 0.604) \text{ other cancer} + (7.482, -0.148) \text{ intestinal obstruction} - (1.422, -0.121) \text{ colorectal} - (1.402, 1.004) \text{ weight loss} + (2.072, -2.013) \text{ diarrhoea} + (10.915, 1.613) \text{ blood stool} + (6.392, -0.499) \text{ anaemia} + (8.790, 0.755) \text{ abdominal}.$

3.3. Fuzzy Linear Regression (Ni, 2005)
The model has been employed by this study to detect tumor size among patients. Twenty five of predictor variables had been used against tumour size (mm).
The evaluation of fuzzy linear regression by Ni (2005) was also mean square error and root mean square error. Mean square error for this model was 110.844 and root mean square error was 10.528.

The estimated fuzzy linear regression parameter by (Ni, 2005) for colorectal cancer of patients is as follow:

\[
\hat{Y} = -8.50\times10^{-14} + (6.417, 6.144\times10^{-14}) \text{gender} - (0.008, 5.149\times10^{-16}) \text{age} + (5.545, -4.767\times10^{-14}) \text{ethnic} + (8.795, -6.025\times10^{-14}) \text{idc10} + (2.675, -5.659\times10^{-15}) \text{TNM Staging} + (16.991, -2.468\times10^{-14}) \text{family history} + (0.320, 1.732\times10^{-14}) \text{diabetes mellitus} - (2.810, -1.818\times10^{-14}) \text{Crohn’s disease} - (2.394, 4.363\times10^{-15}) \text{ulcerative colitis} + (3.291, -7.960\times10^{-15}) \text{polyp} - (3.979, -2.942\times10^{-15}) \text{history of cancer} + (1.127, 1.369\times10^{-14}) \text{endometrial} + (8.001, 2.759\times10^{-15}) \text{gastric} + (6.299, -1.873\times10^{-14}) \text{small bowel} + (1.973, -3.666\times10^{-16}) \text{hepatobiliary} - (2.859, 1.236\times10^{-14}) \text{urinary tract} - (11.855, 9.250\times10^{-15}) \text{ovarian} + (0.498, -4.120\times10^{-15}) \text{other cancer} + (7.428, -2.677\times10^{-14}) \text{intestinal obstruction} - (1.422, 1.850\times10^{-14}) \text{colo} - (1.402, 1.746\times10^{-14}) \text{weight loss} + (2.072, -2.706\times10^{-14}) \text{diarrhoea} + (10.915, 1.990\times10^{-14}) \text{blood stool} + (6.392, -1.123\times10^{-14}) \text{anaemia} + (8.790, -9.496\times10^{-16}) \text{abdominal}.
\]

3.4. Extended Fuzzy Linear Regression By Benchmarking Models Under Fuzziness (Chung, 2012)

Twenty five of predictor variables had been used against tumour size (mm). Extended fuzzy linear regression benchmarking models under fuzzy environment can be developed by means of fuzzy linear regression among variables. A normalization of coefficient is applied to this model to get better results.

There were twenty five independent variables analysed and a new coefficient produced and mean square error for the benchmarking under fuzziness model colorectal cancer was 802.3071. Furthermore, root mean square error value was 28.32503.

The estimated parameter of fuzzy linear regression benchmarking under fuzziness model for colorectal cancer of patients is as follow:

\[
\hat{Y} = 76.056 + 715.266 \text{gender} - 3.806 \text{age} + 615.218 \text{ethnic} + 1029.533 \text{idc10} + 85.430 \text{TNM Staging} - 9959.389 \text{family history} + 326.564 \text{diabetes mellitus} + 1962.528 \text{Crohn’s disease} + 557.707 \text{ulcerative colitis} - 572.597 \text{polyp} + 2106.910 \text{history of cancer} - 325.925 \text{endometrial} - 2405.275 \text{gastric} - 1244.425 \text{small bowel} - 402.856 \text{hepatobiliary} + 960.352 \text{urinary tract} + 5241.343 \text{ovarian} + 1172.459 \text{other cancer} - 1589.287 \text{intestinal obstruction} + 442.145 \text{colo} + 1106.008 \text{weight loss} + 277.794 \text{diarrhoea} - 2844.003 \text{blood stool} - 1096.827 \text{anaemia} - 1429.962 \text{abdominal}.
\]
3.5. Summary of Results

Table 1: Summary of models

<table>
<thead>
<tr>
<th>Models Of Linear Regression</th>
<th>MSE</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple Linear Regression</td>
<td>129.558</td>
<td>11.382</td>
</tr>
<tr>
<td>Fuzzy Linear Regression (Tanaka 1982)</td>
<td>142.794</td>
<td>11.950</td>
</tr>
<tr>
<td>Fuzzy Linear Regression (Ni, 2005)</td>
<td>110.844</td>
<td>10.528</td>
</tr>
<tr>
<td>Extended FLR By Benchmarking Models Under Fuzziness (Chung, 2012)</td>
<td>802.3071</td>
<td>28.325</td>
</tr>
</tbody>
</table>

Based on Table 1, the present study proves that fuzzy linear regression (Ni, 2005) model is the best model to predict tumor size faced by patients in general hospital around Kuala Lumpur. Furthermore, analysis of error was based on two methods, mean square error and root mean square error. It was also used as deciding criteria between multiple linear regression, fuzzy linear regression (Tanaka 1982), fuzzy linear regression (Ni, 2005) and extended fuzzy linear regression by benchmarking models under fuzziness (Chung, 2012) model. In conclusion, fuzzy linear regression (Ni, 2005) is the best model with the smallest value of error which are 110.844 (MSE value) and 10.52 (RMSE value).

4. Conclusions

This study applied four models of linear regression which were multiple linear regression, fuzzy linear regression (Tanaka 1982), fuzzy linear regression (Ni, 2005) and extended fuzzy linear regression by benchmarking models under fuzziness (Chung, 2012) model. The best model to predict tumor size were based on the result of mean square error and root mean square error. Fuzzy linear regression (Ni, 2005) model proved to be the best model in predicting stages of colorectal cancer patients in General Hospital around Kuala Lumpur.

It can be concluded that the tumour size is directly proportional to gender, ethnic, icd10, TNM staging, family history, diabetes mellitus, polyp, endometrial, gastric, small bowel, hepatobiliary, other cancer, intestinal obstruction, diarrhoea, blood stool, anaemia and abdominal. Furthermore, the tumour size is inversely proportional to age, crohn’s disease, ulcerative colitis, history of cancer, urinary tract, ovarian, colorectal and weight loss. The vagueness of tumour size can be represented as the fuzziness of the constant parameter gender, ethnic, age, icd10,
Use of fuzzy linear regression models for tumor size

TNM staging, family history of colon cancer, diabetes mellitus, Crohn's disease, ulcerative colitis, polyp, history of cancer(s), endometrial, gastric, small bowel, hepatobiliary, urinary tract, ovarian, other cancer, intestinal obstruction, colorectal, weight loss, diarrhoea, anaemia, blood stool and abdominal pain. Mean square error value is 110.844 and root mean square error is 10.528.

Acknowledgements. The research work is supported by ERGS (Exploratory Research Grant Scheme) grant (vot E020), Ministry of Higher Education, Malaysia.

References


http://dx.doi.org/10.1109/i4ct.2014.6914150

http://dx.doi.org/10.1158/1055-9965.epi-09-0090


http://dx.doi.org/10.1080/01969727308546046

http://dx.doi.org/10.1109/tsmc.1982.4308925


Received: February 7, 2015; Published: April 3, 2015